"Express Mail" mailing label number EV 164034203 US

Date of Deposit: January 15, 2004

Attorney Docket No.: 15155US01

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

CLASSIFICATION OF SPEECH AND MUSIC USING LINEAR PREDICTIVE CODING COEFFICIENTS

FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0001] [Not Applicable]

[MICROFICHE/COPYRIGHT REFERENCE]

[0002] [Not Applicable]

BACKGROUND OF THE INVENTION

[0003] Human beings, with normal hearing, are often able to distinguish sounds from about 20 Hz, such as the lowest note on a large pipe organ, to 20,000 Hz, such as the high shrill of a dog whistle. Human speech, on the other hand, ranges from 300 Hz to 4,000 Hz.

[0004] Music may be produced by playing musical instruments. Musical instruments often produce sounds that lie outside the range of human speech, and in many instances, produce sounds (overtones, etc.) that lie outside the range of human hearing.

[0005] An audio communication can comprise either music, speech or both. However, conventional equipment processes audio communication signals comprising only speech in a similar manner as communication signals comprising music.

[0006] Further limitations and disadvantages of conventional and traditional approaches will become apparent to one of skill in the art, through comparison of such systems with embodiments presented in the remainder of the present application with references to the drawings.

SUMMARY OF THE INVENTION

[0007] Presented herein are systems and methods for classifying an audio signal.

[0008] In one embodiment of the present invention, there is presented a method for classifying an audio signal. The method comprises calculating a plurality of linear prediction coefficients for a portion of the audio signal; inverse filtering the portion of the audio signal with the plurality of linear prediction coefficients filter, thereby resulting in a residual signal; measuring the energy of the residual signal; and comparing the residual energy to a threshold.

[0009] In another embodiment, the method further comprises classifying the portion of the audio signal as music, if the residual energy exceeds the threshold; and classifying the portion of the audio signal as speech, if the threshold exceeds the residual energy.

[00010] In another embodiment, the portion of the audio signal comprises a frame.

[00011] In another embodiment, the method further comprises decimating the frame, thereby causing the frame to comprise a predetermined number of samples.

[00012] In another embodiment, the method further comprises spectrally flattening the portion of the audio signal.

[00013]In another embodiment, there is presented a method for classifying an audio signal.

[00014]The method comprises taking a discrete Fourier transformation of a portion of the audio signal for a plurality of frequencies; calculating a plurality of linear prediction coefficients (LPC) for the portion of the signal; measuring an inverse filter response for said plurality of frequencies with said plurality of linear prediction coefficients (LPC); measuring a mean squared error between the discrete Fourier transformation of the portion of the audio signal for the plurality of frequencies and the inverse filter response; and comparing the means squared error to a threshold.

[00015]In another embodiment, the method further comprises classifying the portion of the audio signal as music, if the mean squared error exceeds the threshold; and classifying the portion of the audio signal as speech, if the threshold exceeds the means squared error energy.

[00016]In another embodiment, the portion of the audio signal comprises a frame.

[00017]In another embodiment, the method further comprises decimating the frame, thereby causing the frame to comprise a predetermined number of samples.

[00018]In another embodiment, the method further comprises spectrally flattening the portion of the audio signal.

[00019]In another embodiment, there is presented a system for classifying an audio signal. The system comprises a first circuit, an inverse filter, a second circuit, and a third circuit. The first circuit calculates a plurality of linear prediction coefficients for a portion of the audio signal. The inverse filter inverse filters the portion of the audio signal with the plurality of linear prediction coefficients, thereby resulting in a residual signal. The second circuit measures the energy of the residual signal. The third circuit compares the residual energy to a threshold.

[00020]In another embodiment, the system further comprises logic for classifying the portion of the audio signal as music, if the residual energy exceeds the threshold, and classifying the portion of the audio signal as speech, if the threshold exceeds the residual energy value.

[00021]In another embodiment, the portion of the audio signal comprises a frame.

[00022]In another embodiment, the system further comprises a decimator for decimating the frame, thereby causing the frame to comprise a predetermined number of samples.

[00023]In another embodiment, the system further comprises a pre-emphasis filter for spectrally flattening the portion of the audio signal.

[00024]In another embodiment, there is presented a system for classifying an audio signal. The system comprises a first circuit, a second circuit, an inverse filter, a third

circuit, and a fourth circuit. The first circuit takes a discrete Fourier transformation of a portion of the audio signal for a plurality of frequencies. The second circuit calculates a plurality of linear prediction coefficients (LPC) for the same portion of the signal. The inverse filter measures an inverse filter response for said plurality of frequencies with said plurality of linear prediction coefficients (LPC). The third circuit measures a mean squared error between the discrete Fourier transformation of the portion of the audio signal for the plurality of frequencies and the inverse filter response. The fourth circuit compares the means squared error to a threshold.

[00025] In another embodiment, the system further comprises logic for classifying the portion of the audio signal as music, if the mean squared error exceeds the threshold and classifying the portion of the audio signal as speech, if the threshold exceeds the means squared error energy. In another embodiment, the portion of the audio signal comprises a frame.

[00026] In another embodiment, the system further comprises a decimator for decimating the frame, thereby causing the frame to comprise a predetermined number of samples.

[00027] In another embodiment, the system further comprises a pre-emphasis filter for spectrally flattening the portion of the audio signal.

[00028]These and other advantages and novel features of the present invention, as well as details of an illustrated example embodiment thereof, will be more fully understood from the following description and drawings.

Brief Description of the Drawings

[00029] **FIGURE 1** is a flow diagram for classifying a digital audio signal as speech or music in accordance with an embodiment of the present invention;

[00030] **FIGURE 2** is a flow diagram for classifying a digital audio signal as speech or music in accordance with an alternative embodiment of the present invention;

[00031] **FIGURE 3** is a system for classifying a digital audio signal as speech or music in accordance with an embodiment of the present invention;

[00032] **FIGURE 4** is a system for classifying a digital audio signal as speech or music in accordance with an alternative embodiment of the present invention;

[00033] **FIGURE 5** is a block diagram illustrating a system for converting, classifying, encoding, and packetizing an audio communication according to an embodiment of the present invention;

[00034] **FIGURE 6** is a block diagram illustrating encoding of an exemplary audio signal according to an embodiment of the present invention; and

[00035] **FIGURE 7** is a block diagram illustrating an exemplary audio decoder according to an embodiment of the present invention.

[00036] Referring now to **FIGURE 1**, there is illustrated a flow diagram for classifying whether a digital audio signal is speech or music. At 105, the digital audio signal is divided into a set of frames. The frames comprise a fixed number of digital audio samples from the digital audio signal. Additionally, frames can be processed in a number of ways, such as by a decimator, pre-emphasis filter, or a windowing function, to name a few.

[00037] At 110, a finite number of Linear Prediction coefficients (LPC) are calculated for each frame. In general, the inherent limitations of the human vocal tract allow a speech signal spectrum to be shaped by fewer LPC coefficients than a music signal. Accordingly, at 115 the inverse filter response of the frame to an inverse filter according to the LPC coefficients (the residual signal) calculated during 110 is taken and the residual energy is measured at 117. The residual energy of the filter response is compared at 120 to an energy threshold.

[00038] If the residual energy exceeds the threshold, at 120, the frame is classified (125) as music. If the residual energy does not exceed the threshold at 120, the frame is classified (130) as speech.

[00039] Referring now to **FIGURE 2**, there is illustrated a flow diagram for classifying a digital audio signal as speech or music in accordance with an alternative embodiment of the present invention. At 55, the digital audio signal is divided into a set of frames. The frames comprise a fixed number of digital audio samples from the digital audio signal. Additionally, frames can be processed

in a number of ways, such as by a decimator, pre-emphasis filter, or a windowing function, to name a few.

[00040] At 60, the Discrete Fourier Transformation (DFT) is taken for a frame. At 65, the LPC coefficients are determined. At 70, the LPC inverse filter response is taken and measured for the DFT frequencies. At 75, the mean squared error is calculated and compared to a threshold at 80.

[00041] If the means squared error exceeds the threshold, at 230, the frame is classified (85) as music. If the mean squared error does not exceed the threshold at 80, the frame is classified (90) as speech.

[00042] Referring now to **FIGURE 3**, there is illustrated a block diagram describing an exemplary system for classifying a digital audio input signal 105 as speech or music. The digital audio input signal 105 can be from any real time audio source or recorded data from any other medium.

[00043] A decimator filter 110 receives the digital audio input signal 105 and divides the digital audio input signal 105 into smaller blocks containing a finite number of audio samples called a frame. The frame size depends upon the sampling rate of the digital audio input signal 105, because the decimator filter 110 provides a fixed number of samples per frame, and a fixed number of frames per second. For example, if the digital audio input signal 105 is sampled at 48000 samples/second, and the decimator filter 110 provides 50 frames comprising 160 samples, per second, the frame size can be set at 960 samples per frame, and the decimation factor set at six. The decimator filter 110 can

be an adaptive filter that decimates the given audio samples appropriately in such a way that the output of the decimator filter 110 is at a fixed rate.

[00044] A pre-emphasis filter 115 receives the output 112 of the decimator filter 110. The pre-emphasis filter 115 may be a first-order finite impulse response (FIR) filter that spectrally flattens the output 112 of the decimator filter 110. The pre-emphasis filter can have the transfer function:

$$H(z) = 1/(1 + a_{pre}z^{-1})$$

[00045] The pre-emphasis factor a_{pre} can be approximately 15/16. The pre-emphasis filter 115 removes the DC component of the audio signal and helps in improving the estimation of Linear Prediction Coefficients (LPC) from auto-correlation values.

[00046] A windowing function 120 receives the output 117 of the pre-emphasis filter 115. The windowing function 120 can comprise any one of a number of different windowing standards, such as, Hamming, Hanning, Blackman, or Kaiser windows. The individual frames are windowed to minimize the signal discontinuities at the borders of each frame. If the window is defined as w[n], $0 < n < N-1$, then the windowed signal is s[n] = w[n]*u[n], where u[n] is the initial input data before windowing.

[00047] An auto-correlation coefficients computation function 125 receives the output of the windowing function 120. In an exemplary case, the windowed frame S comprises 160 samples, where S = (s(0), s(1),...s(159)). In a case where the frame comprises 160 samples, a 10^{th} order LPC

coding is sufficient to model the spectrum if S is a speech signal. The signal s[n] is related to the innovation u[n] signal[The error signal between the actual signal and signal predicted using this 10^{th} order LPC coefficients] through the linear difference equation:

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n)$$

[00048] These 10 LPC coefficients are chosen to minimize the energy of the innovation signal u[n]:

$$f = \sum_{n=0}^{159} u^2(n)$$

[00049] The foregoing can be determined by taking the derivative with respect to a_i, and setting the derivative to zero as shown below:

$$df/da_1 = 0$$

$$df/da_2 = 0$$

$$\cdots\cdots$$

$$df/da_{10} = 0$$

[00050] The above can be simplified to get 10 linear equations with 10 unknowns, the unknowns being the LPC coefficients. The 10 equations can be represented by the martix below:

$$\begin{bmatrix} R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) & R(9) \\ R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) \\ R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) \\ R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) \\ R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) \\ R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) \\ R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) \\ R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) \\ R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) \\ R(9) & R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \end{bmatrix} = \begin{bmatrix} -R(1) \\ -R(2) \\ -R(3) \\ -R(4) \\ -R(5) \\ -R(6) \\ -R(7) \\ -R(8) \\ -R(9) \\ -R(10) \end{bmatrix}$$

[00051] Where

$$\begin{aligned} R(k) &= \sum_{n=0}^{159-k} s(n)s(n+k) \\ &= \text{autocorrelation of } s(n) \end{aligned}$$

[00052] The auto-correlation coefficients computation function 125 provides the auto-correlation coefficients R(k) to the LPC coefficients computation function 130. The LPC coefficients are determined by calculating $a_1 \ldots a_{10}$ from the above matrix. The above matrix can be solved using the Gaussian elimination method, matrix inversion, or Levinson-Durbin recursion. However, since the above matrix is a Toeplitz matrix (symmetrical & diagonals equal), the standard Levinson-Durban recursion is advantageous.

[00053] The LPC coefficients are provided from the LPC Coefficients Computation function 130 to an Inverse LPC Analysis Filter 135. The LPC analysis filter filters the input data s[n]. Since a 10^{th} order LPC filter response very closely represents the gross shape of a given input speech signal spectrum for a frame comprising 160 samples, if the given audio signal s[n] represents speech, the residual energy will be very small in comparison to the input audio signal energy. In contrast, if the given audio signal s[n] represents music, the residual energy will be significant in comparison to the input audio signal energy.

$$\text{Input signal energy} = \sum_{n=0}^{n=159} s^2[n]$$

$$\text{Residual signal energy} = \sum_{n=0}^{n=159} r^2[n]$$

[00054] In some cases, it may not be easy to decide clearly about speech or music for a specific frame since the energy ratio value may be very close to the threshold value. In such cases, the decision may be delayed for few frames and final decision for all the frames is taken jointly depending upon the majority of the frame decisions. Each frame decision (i.e. speech or music) is taken the same way by comparing the ratio of the residual signal energy to input signal energy against the ENERGY_THRESHOLD (0.15) value for all the frames but final decision for all the audio frames is taken at the end only depending upon the majority of all the decisions.

[00055] If the ratio of residual signal energy to input signal energy is very close to the ENERGY_THRESHOLD value then decision is delayed for that frame and the same algorithm is applied to the next two or four consecutive frames depending upon the energy ratio value. Once, the individual decision is taken for all the three/five frames. With majority logic 140, whatever decisions (either speech or music) are more for all the frames, that same decision is applied to all three/five frames together.

[00056] Referring now to **FIGURE 4**, there is illustrated a block diagram of a system for classifying an input digital audio signal as music or speech in accordance with an alternative embodiment of the present invention. The Fourier transform of the given input signal s[n] is taken for a finite number of points and the magnitude of all 512

uniformly spaced frequency values are computed by a DFT function 145. The LPC filter response also at all those same 512 frequency values is sampled and the magnitude of all those 512 frequency values are computed by LPC filter sampling function 150.

[00057] With the frequency magnitudes vector for all 512 frequencies from both DFT function 145 and LPC filter sampling function 150, the mean squared error value for all the frequencies is computed by a means squared error computation function 155. Once the mean squared value is computed, the value is compared against a SQUARED_ERROR_THRESHOLD value. If the value is below that threshold value, it will be declared a speech frame, otherwise it will be declared a music frame.

$$\text{Mean squared error} = \frac{1}{512}\sum_{f=0}^{f=511}[S(f)-H(f)]^2$$

[00058] In some cases, it may not be easy to decide clearly about speech or music for a specific frame since the mean squared error value may be very close to the threshold value. In such cases, the decision may be delayed for few frames and final decision for all the frames is taken jointly depending upon the majority logic 140. It means that the frame decision (i.e. speech or music) is taken the same way by comparing the mean squared error value against the SQUARED_ERROR_THRESHOLD value for all the frames.

[00059] If the ratio of mean squared error value is very close to SQUARED_ERROR_THRESHOLD value then decision is delayed for that frame and the same algorithm is applied to

the next two or four consecutive frames depending upon the mean squared error value. The individual decision is taken for all the three/five frames one time.

[00060] **FIGURE 5** is a block diagram illustrating a system 800B for converting, classifying, encoding, and packetizing an audio communication according to an embodiment of the present invention. The system 800B receives an audio communication 810B, wherein the audio communication 810B may be either an analog signal 801B or a digital signal 803B. The audio communication 810B may proceed directly to speech/music classification apparatus 866B as an analog signal 801B at junction 863B. Alternatively, the audio signal 810B may be passed through analog to digital converter 805B for conversion to a digital signal 803B that is provided via junction 797 to the speech/music classification apparatus 866B. After conversion from analog to digital, the digital signal 803B may be passed to MPEG encoder 825B. The circumstances of the audio signal processing at the MPEG encoder 852B will be described below.

[00061] The audio signal may arrive at the speech/music classifying apparatus 866B at input 820B. The signal is then passed to mathematical processor 830B. After the mathematical processing has been completed and the ratio is determined, the ratio is passed to comparator 860B. Comparator 860B is adapted to compare the calculated ratio to the threshold value. The threshold value may be pre-set by a user, or the comparator 860B may determine (learn) the threshold value through trial and error. If the ratio is greater than the threshold value, then the output from the speech/music classifying apparatus 866B is that the audio

signal is determined to be music. However, if the ratio is less than the threshold value, then the output from the classifying apparatus 866B is that the audio signal is speech.

[00062] The signal may then be passed to either encoder 825B or alternatively to packetization engine 835B via junction 895B. In one embodiment, encoder 825b comprises an MPEG encoder. The encoder 825B converts the digital signal 803B to an audio elementary stream (AES), AES encoding the digital signal 803B in accordance with the MPEG standard, for example. When the AES is directed to the packetization engine 835B, the AES is packetized into a packetized audio elementary stream comprising packets 855B. Each packet comprising a portion of the AES and may also comprises a flag 875B. The flag 875B may indicate that the portion of the AES in the packet is speech or music depending upon the state of the flag 875B, i.e., whether the flag is turned on or off.

[00063] **FIGURE 6** is a block diagram 800C illustrating encoding of an exemplary audio signal A(t) 810C by the encoder 825B according to an embodiment of the present invention. The audio signal 810C is sampled and the samples are grouped into frames 820C ($F_0 \ldots F_n$) of 1024 samples, e.g., ($F_x(0) \ldots F_x(1023)$). The frames 820C ($F_0 \ldots F_n$) are grouped into windows 830C ($W_0 \ldots W_n$) that comprise 2048 samples or two frames, e.g., ($W_x(0) \ldots W_x(2047)$). However, each window 830C W_x has a 50% overlap with the previous window 830C W_{x-1}.

[00064] Accordingly, the first 1024 samples of a window 830C W_x are the same as the last 1024 samples of the previous window 830C W_{x-1}. A window function w(t) is applied to each

window 830C ($W_0 \ldots W_n$), resulting in sets ($wW_0 \ldots wW_n$) of 2048 windowed samples 840C, e.g., ($wW_x(0) \ldots wW_x(2047)$). The modified discrete cosine transformation (MDCT) is applied to each set ($wW_0 \ldots wW_n$) of windowed samples 840C ($wW_x(0) \ldots wW_x(2047)$), resulting sets ($MDCT_0 \ldots MDCT_n$) of 1024 frequency coefficients 850C, e.g., ($MDCT_x(0) \ldots MDCT_x(1023)$).

[00065] The encoder 825B receives the output of the speech/music classification 866B apparatus. Based upon the output of the speech/music classification apparatus 866B, the encoder 825B can take any number of actions with respect to the MDCT coefficients. For example, where the output indicates that the content associated with the audio signal 810C is speech, the encoder 825B can either discard or quantize with fewer bits the MDCT coefficients associated with frequencies outside the range of human speech, i.e., exceeding 4 KHz. Where the output indicates that the content associated with the audio signal 810C is music, the MPEG 825B can quantize the MDCT coefficients associated with frequencies outside the range of human speech.

[00066] The sets of frequency coefficients 850C ($MDCT_0 \ldots MDCT_n$) are then quantized and coded for transmission, forming what is known as an audio elementary stream (AES). The AES can be multiplexed with other AESs. The multiplexed signal, known as the Audio Transport Stream (Audio TS) can then be stored and/or transported for playback on a playback device. The playback device can either be local or remotely located.

[00067] Where the playback device is remotely located, the multiplexed signal is transported over a communication medium, such as the Internet. During playback, the Audio

TS is de-multiplexed, resulting in the constituent AES signals. The constituent AES signals are then decoded, resulting in the audio signal.

[00068] Alternatively, the frequency coefficients $MDCT_0$ . . . $MDCT_n$ may be packetized by the packetization engine of **FIGURE 6**. In an audio signal, each frame may comprise frequency coefficients 850C ($MDCT_0$...$MDCT_{1023}$). Sub-frame contents may correspond to a particular range of audio frequencies.

[00069] **FIGURE 7** is a block diagram illustrating an exemplary audio decoder 900 according to an embodiment of the present invention. Referring now to **Figure 7**, once the frame synchronization is found and delivered from signal processor 901, the advanced audio coding (AAC) bit stream 903 is de-multiplexed by a bit stream de-multiplexer 905. This includes Huffman decoding 916, scale factor decoding 915, and decoding of side information used in tools such as mono/stereo 920, intensity stereo 925, TNS 930, and the filter bank 935.

[00070] The sets of frequency coefficients 850C ($MDCT_0$...$MDCT_n$) are decoded and copied to an output buffer in a sample fashion. After Huffman decoding 916, an inverse quantizer 940 inverse quantizes each set of frequency coefficients 850C ($MDCT_0$...$MDCT_n$) by a 4/3-power nonlinearity. The scale factors 915 are then used to scale sets of frequency coefficients 850C ($MDCT_0$...$MDCT_n$) by the quantizer step size.

[00071] Additionally, tools including the mono/stereo 920, prediction 923, intensity stereo coupling 925, TNS 930, and filter bank 935 can apply further functions to the sets of

frequency coefficients 850C ($MDCT_0 \ldots MDCT_n$). The gain control 950 transforms the frequency coefficients 850C ($MDCT_0 \ldots MDCT_n$) into the time domain signal A(t). The gain control 950 transforms the frequency coefficients 850C by application of the Inverse MDCT (IMDCT), the inverse window function, window overlap, and window adding. The gain control 950 also looks at the flag 875B. The flag 875B is a bit that may be either on or off, i.e., having binary digital value of 1 or zero, respectively. For example, if the bit is on, this indicates that the audio signal is music, and if the bit is off, this indicates that the audio signal is speech, or vice versa.

[00072] If the flag 875B indicates that the audio signal is music the gain control and may then perform the decoding by performing the Inverse MDCT function. The gain control 950 may also report results directly to the audio processing unit 999 for additional processing, playback, or storage. The gain control 950 is adapted to detect at the receiving/decoding end of the audio transmission whether the audio signal is one of music or speech.

[00073] Another music/speech classifier 966, such as the systems disclosed in **FIGURES 3 or 4**, may be provided at the decoder 900, so that in the circumstance where the signal has been received at the decoder 900 without being classified as one of speech or music, the signal may then be classified. The signal may also be passed to an audio processing unit 999 for storage, playback, or further analysis, as desired.

[00074] One embodiment of the present invention may be implemented as a board level product, as a single chip, application specific integrated circuit (ASIC), or with

varying levels integrated on a single chip with other portions of the system as separate components. The degree of integration of the system will primarily be determined by speed and cost considerations. Because of the sophisticated nature of modern processors, it is possible to utilize a commercially available processor, which may be implemented external to an ASIC implementation of the present system. Alternatively, if the processor is available as an ASIC core or logic block, then the commercially available processor can be implemented as part of an ASIC device with various functions implemented as firmware.

[00075] The foregoing description of the exemplary embodiment of the invention has been presented for the purposes of illustration and description. While the invention has been described with reference to certain embodiments, it will be understood by those skilled in the art that various changes may be made and equivalents may be substituted without departing from the scope of the invention. In addition, many modifications may be made to adapt a particular situation or material to the teachings of the invention without departing from its scope. Therefore, it is intended that the invention not be limited to the particular embodiment disclosed, but that the invention will include all embodiments falling within the scope of the appended claims.